High-Performance Throughput Tuning/Measurements

Davide Salomoni & Steffen Luitz

Presented at the PPDG Collaboration Meeting,
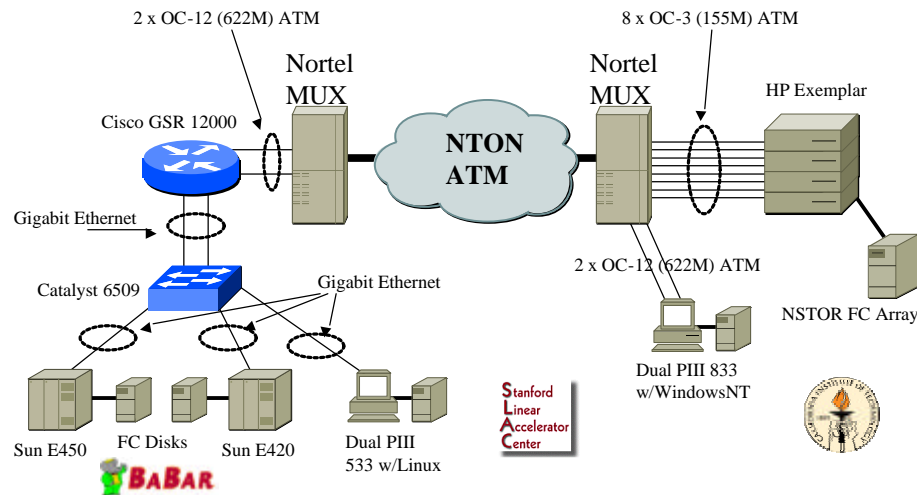*Argonne National Lab, July 13-14, 2000*

http://www.slac.stanford.edu/grp/scs/net/talk/High_Perf_PPDG_Jul2000.ppt

---

The Goals

- In the framework of the Particle Physics Data Grid (PPDG) project, to demonstrate WAN I/O transfer at a rate of 100+ MB/sec
  - Using the NTON test-bed network between SLAC and Caltech
- Also, try to assess actual performance of high-speed NICs
  - For internal capacity planning
  - To validate use as high-speed sniffers?
- Question #1: in the first place, what do we need / how is it possible to achieve 100 MB/s on the LAN?
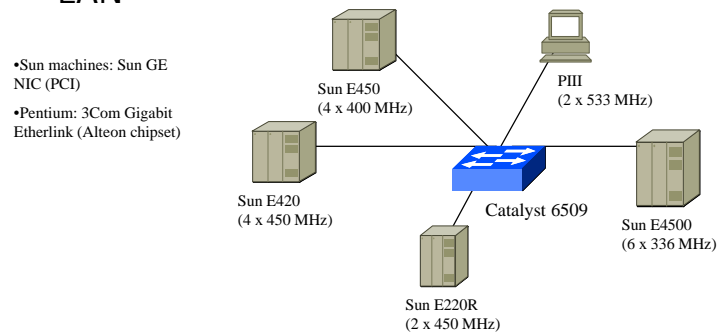- Question #2: what about the WAN?

Connection Details

2 x OC-12 (622M) ATM

8 x OC-3 (155M) ATM

Nortel MUX

Nortel MUX

HP Exemplar

Cisco GSR 12000

NTON ATM

Gigabit Ethernet

Catalyst 6509

Gigabit Ethernet

2 x OC-12 (622M) ATM

NSTOR FC Array

Sun E450   FC Disks   Sun E420

Dual PIII 533 w/Linux

Dual PIII 833 w/WindowsNT

Stanford **L**inear **A**ccelerator **C**enter

BaBar

3

Local Tests

❚ Before using NTON, we set up a separate local test-bed to verify GE NICs performance in the LAN

•Sun machines: Sun GE NIC (PCI)

•Pentium: 3Com Gigabit Etherlink (Alteon chipset)

Sun E450
(4 x 400 MHz)

PIII
(2 x 533 MHz)

Sun E420
(4 x 450 MHz)

Catalyst 6509

Sun E4500
(6 x 336 MHz)

Sun E220R
(2 x 450 MHz)

4

Local Tests: Results

❚ The tool used to check TCP and UDP throughput is *iperf*:
http://dast.nlanr.net/Projects/Iperf/release.html

❚ Performance of Sun Enterprise HW (220, 420/450, 4500) and of Sun GE cards is rather shabby

❚ Solaris (2.6/8) also seems to be not too performing

❚ Linux (especially the latest kernel, 2.3.x) quite outperformed everything else – even with just 2 PIII 533 MHz processors

5

NTON Tests

❚ Still ongoing… but:
  ❚ Current setup has a 6-CPU Sun E4500 w/Solaris and a Dual-Pentium III 533 w/Linux on the SLAC side (with a Cisco 12000 doing the WAN routing), and a Dual-Pentium III 833 w/WindowsNT and 2 Fore Systems OC-12 ATM interfaces on the Caltech side;
  ❚ Using a single OC-12 from Caltech to the SLAC E4500, got ~320 Mbit/s; with two OC-12, got ~600 Mbps; we are working and try to understand how to go higher (large number of retransmits, CPU load quite high on the Caltech side).
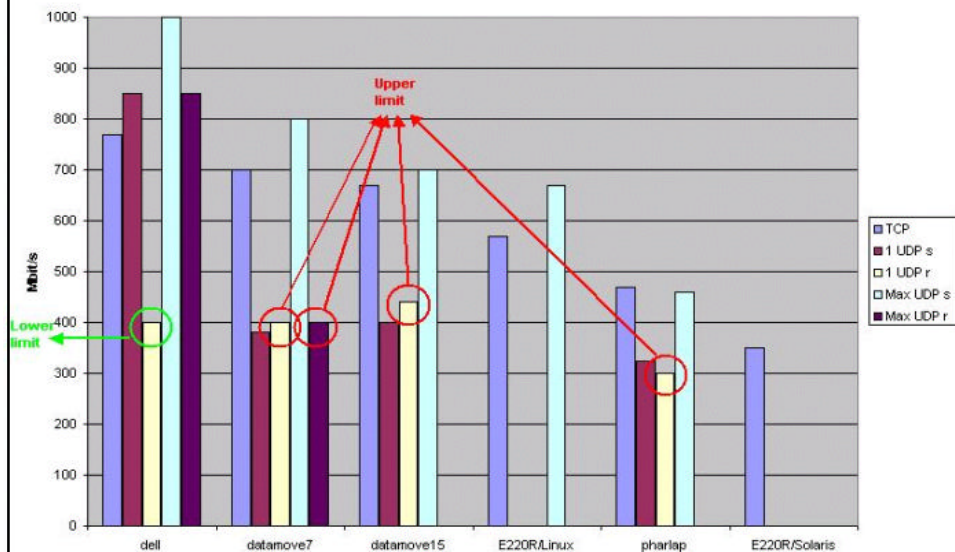
6

LAN TCP thruput from 2*533MHz PIII cpu to 4*400MHz Sun E450
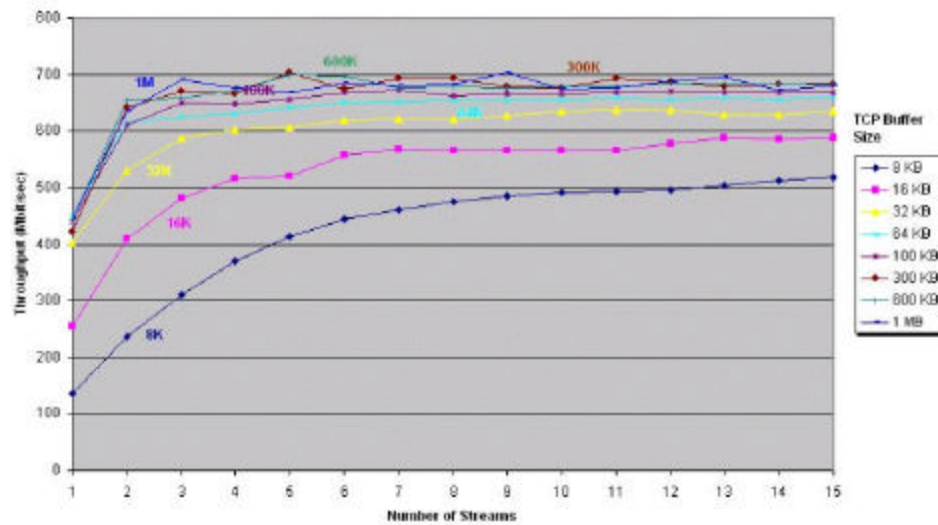

7

Comparing Sun/Solaris vs PIII/Linux

8

4

Results from LAN tests

- Thruput CPU limited (tried with 2 GE interfaces)
  - 2*533MHz PIII TCP thruput < 770Mbps
- For good performance need to tune:
  - Multi-streams important
  - Need window size optimization
    - Window = bandwidth * RTT
- With UDP PIII/2 * 553MHz cpu can drive:
  - 1200Mbps with 2 GE & 100% CPU util. & MTU 1500Bytes
  - 1500Mbps with 2 GE & 50% CPU util. & 9000Byte jumbo frames
  - BUT heavy loss on receiver side after 400Mbps
- Sun/Solaris disappointing vs. Linux

9

WAN TCP thruput
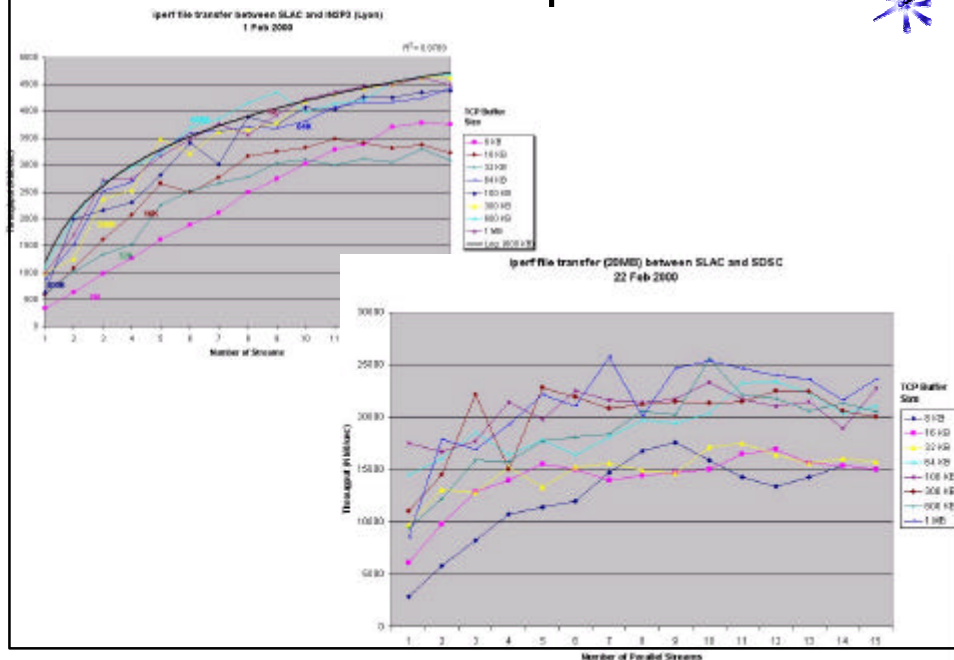
Measured WAN thruput
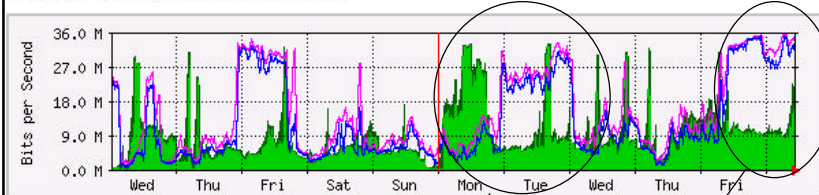
SLAC to	RTT ms.	Window size kBytes	Bottleneck bandwidth Mbps	Meas. TCP thruput Mbps
SLAC	0.225 ms	300 kB	1000M	700M
San Diego	17 ms	64 kB	30 M	30 M
CERN	177 ms	270 kB (theory 600kB)	27 M	25 M
Italy	330 ms	60 kB	1.5 M*	1.5 M
Lyon, Fr	190 ms	140 kB	6 M	4.5 M

11

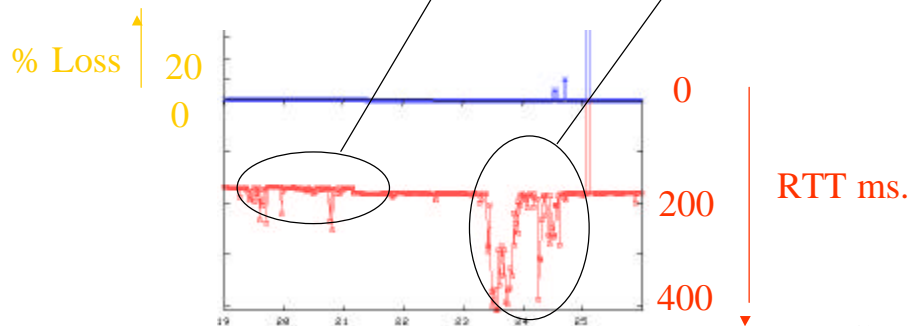# Impact

`Weekly' Graph (30 Minute Average)

Max In:33.5 Mb/s (74.3%)   Average In:8190.4 kb/s (18.2%)   Current In:11.3 Mb/s (25.1%)
Max Out:35.5 Mb/s (78.9%) Average Out: 12.8 Mb/s (28.5%) Current Out:34.1 Mb/s (75.8%)

% Loss   20

0

0

200   RTT ms.

400

12

Compression

▌ Bbftp tool from Gilles Farrache, IN2P3

  ▌ 10 streams SLAC > Lyon got about 38Mbps

    ▎ CPU power is important

      • The best a Sun E4500 with 4 CPUs @336Mhz could do was 13.6Mbps with 5 streams, more streams go slower (e.g. 10 7.4Mbps)

      • 38 Mbps was done on 64 400MHz CPU Sun E10000

      • The compression ratio is typically of the order of 2 to 3 for BaBar Objectivity databases, so the actual data rate gets boosted to 27.2-40.8 Mbps (E4500) or 76-114 Mbps (E10000)

13

WAN conclusions

▌ High FTP performance across WAN links is possible

  ▌ Have got 350GBytes SLAC>IN2P3 in 5 days

▌ OS must support big windows selectable at application level

▌ Loss is important in particular interval between losses

▌ We can saturate links, need reverse QoS

  ▌ Need to notify people today

▌ Compression looks promising

▌ Still struggling with NTON

▌ For more see:

  ▌ http://www.psc.edu/networking/perf_tune.html

  ▌ http://www.slac.stanford.edu/comp/net/wan-mon/slac-cern.html

  ▌ http://www.slac.stanford.edu/~salomoni/perftest/

14